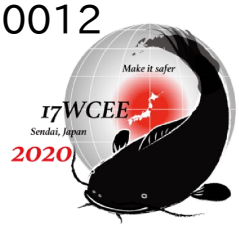


17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 13th to 18th 2020

PERFORMANCE METRICS TO EVALUATE PROBABILISTIC MODELS FOR STRUCTURAL DAMAGE DURING SEISMIC EVENTS

L. Burks[1] and A. Gupta[2]

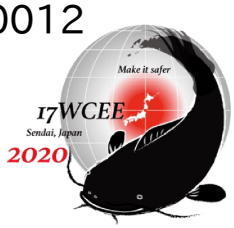
[1] *Seismic Data Scientist; One Concern, Inc.; lynne@oneconcern.com*
[2] *Resilience Engineering Lead; One Concern, Inc.; abhineet@oneconcern.com*

Abstract

Rapid identification of buildings and infrastructure damaged in an earthquake is important to minimize both human and economic losses and to deploy appropriate resources to impacted areas. However, damage to buildings is often rare and highly uncertain. Because of the uncertainty in building vulnerability, models that provide a probabilistic estimate of damage are appropriate. But it can be challenging to evaluate the performance of a probabilistic model against ground-truth observations, and to compare the performance of multiple probabilistic models. This paper investigates metrics that can be used to evaluate probabilistic damage models suitable for earthquake engineering applications.

We discuss the advantages and limitations of the following metrics: (1) expected damage accuracy where damage is estimated based on expectation of the probabilities, (2) threshold damage accuracy where damage is estimated based on a selected threshold on probabilities, (3) rank probability score (RPS) that compares the probabilities directly with observations, and (4) threshold-weighted rank probability score (t-RPS) that improves over the RPS by weighing different regions of the probability distribution based on their importance. We present a simulation study and a case study to demonstrate the discussed advantages and limitations of each metric. The case study uses historical earthquake damage data from the 2014 South Napa earthquake and seismic fragility functions from Hazus to predict probabilistic damage. We demonstrate the utility of the proposed metrics by comparing the performance of Hazus fragility functions to a simple baseline model for the 2014 South Napa earthquake. The expected and threshold damage accuracy metrics are found to not be suitable for the evaluation of probabilistic models, particularly for imbalanced datasets and rare events like severe damage from earthquakes. The RPS and t-RPS metrics are appropriate for this application, as long as the damage data is known to be a representative sample of the overall damage distribution.

Keywords: scoring rules; model evaluation; probabilistic damage prediction; forecasting models



1. Introduction

Rapid identification of buildings and infrastructure damaged in an earthquake is important to minimize both human and economic losses and to deploy appropriate resources to impacted areas. However, damage to buildings is often rare and highly uncertain. Therefore, models that provide a probabilistic estimate of damage are appropriate. But it can be challenging to compare model predictions to ground truth observations when the model predictions are probabilistic and ground truth observations are not. It can also be challenging to compare the performance of one probabilistic model to another.

There are both aleatory and epistemic uncertainties in building and infrastructure damage from earthquakes. Therefore, traditional classification metrics like accuracy, precision, recall, and F1-score are not suitable [1]. These metrics force the model to classify a building into one damage state rather than taking the entire predicted damage distribution into account. In particular, these metrics do not account for the shape of the predicted distribution nor for ordinality in the multiclass damage estimation.

In this paper, we present four metrics that can be used to evaluate probabilistic predictions of damage: (1) expected damage accuracy, (2) threshold damage accuracy, (3) rank probability score (RPS), and (4) threshold-weighted rank probability score (t-RPS). The RPS is an example of a scoring rule commonly used in forecasting applications like weather forecasting [2]. Another common metric used to compare one probability distribution to another is the Kullback-Leibler divergence (KLD) [3]. However, the KLD is not appropriate for this application because we must compare a predicted distribution to an observation. The KLD is only defined for cases in which any zero in the reference distribution is also zero in the predicted distribution. However, in the case of observed damage data, the probability will be 1 for the observation and 0 for all other damage states, making the KLD undefined for any predicted distribution that has non-zero probabilities for any damage state but the observed one.

The rest of this paper is organized as follows: Section 2 defines each of the four proposed metrics, Section 3 presents a simulation study in order to understand their behavior, and Section 4 presents a case study of the metrics applied to actual damage data from the 2014 earthquake in Napa, California, USA.

2. Metrics Overview

In this section, we introduce four metrics for the evaluation of probabilistic models: (1) expected damage accuracy, (2) threshold damage accuracy, (3) rank probability score (RPS), and (4) threshold-weighted rank probability score (t-RPS). In order to describe these metrics, we consider two example models that predict the probability of a building being in one of four damage states after an earthquake. Suppose an earthquake occurs and the building is observed to be in damage state 2 and the model predictions are as shown in Table 1.

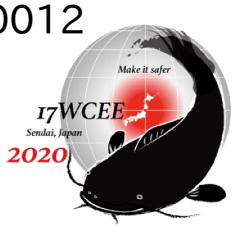
Table 1 – Predicted probabilities of damage for a building from two example models. P[DS=*i*] represents the probability of being in damage state *i*.

Model	P[DS=0]	P[DS=1]	P[DS=2]	P[DS=3]
1	0.05	0.5	0.4	0.05
2	0.2	0.3	0.3	0.2

2.1 Expected Damage Accuracy

Expected damage accuracy is used to evaluate models for ordinal classification [4]. For earthquake damage models, the damage state is estimated by rounding the expectation of the predicted probability. The predicted damage state is then compared to the observed damage state. The predicted damage state is defined as

$$DS = Round\left(\sum_{k=1}^{K} ds_k p_k\right) \quad (1)$$



where K is the number of damage states, ds_k is damage state k, and p_k is the probability of being in damage state k, and *Round* is the rounding operator that returns the closest integer.

For example, consider the case defined in Table 1. For model 1, the estimated damage state is $Round(0 \cdot 0.05 + 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.05) = 1$, and for model 2, the estimated damage state is $Round(0 \cdot 0.2 + 1 \cdot 0.3 + 2 \cdot 0.3 + 3 \cdot 0.2) = 2$. Because the observed damage state is 2, model 1 has an expected damage accuracy of 0% and model 2 has an expected damage accuracy of 100%. The expected damage accuracy is computed for each building then averaged over the entire dataset.

This metric is intuitive and simple to communicate. However, there are several shortcomings of this metric for probabilistic models. The predicted damage states will be concentrated near the mean of the distribution instead of following the true distribution. For example, for a true uniform distribution sampled between 1 and 10, the expectation of the distribution (and thus the predicted label) will be predominantly in the middle of the range, i.e., 5.

Also, observations of the higher damage states are rare, so a well calibrated model will predict a small probability of their occurrence, resulting in those damage states never being predicted. In this case of imbalanced labels, a model can achieve good accuracy by simply always predicting the most common damage state. One way to account for this is to use the class balanced accuracy [5], which computes the average expected accuracy for each damage state and then takes the average over the damage states.

2.2 Threshold Damage Accuracy

Threshold damage accuracy has been used to evaluate weather forecasts for intense events [6]. For earthquake damage prediction, damage is estimated as the highest damage state at which the reverse of the cumulative distribution function (CDF) exceeds a given threshold. The reverse CDF of the prediction is defined as

$$Y_k = P(y \geq y_k) \tag{2}$$

where k is the damage state. For example, the reverse CDF for model 1 in Table 1 is $Y = [1.0, 0.95, 0.45, 0.05]$, and the reverse CDF for model 2 is $Y = [1.0, 0.8, 0.5, 0.2]$.

Then suppose we define a damage threshold of 0.5. For model 1, the predicted damage state is 1 because that is the highest damage state for which $Y_k \geq 0.5$. For model 2, the predicted damage state is 2 because $Y_2 \geq 0.5$. Once a damage state is predicted, the threshold damage accuracy is computed the same way as expected damage accuracy: the predicted damage state is compared to the observed. Therefore, in this example where the observed damage is 2, the threshold damage accuracy for model 1 is 0% and for model 2 is 100%.

The threshold damage accuracy allows for the prediction of higher damage states with the selection of a low threshold. However, selection of too low a threshold will lead to over-prediction of damage. The probability of occurrence of high damage states is typically so low that threshold accuracy will not be able to capture it well. This metric has similar shortcomings to the expected damage accuracy. The predicted damage state will be biased to occur near the quantile selected as the threshold. For example, in a uniform distribution between 1 and 10, 5 will be the most likely label with a threshold of 0.5, and 7 with a threshold of 0.3.

2.3 Rank Probability Score (RPS)

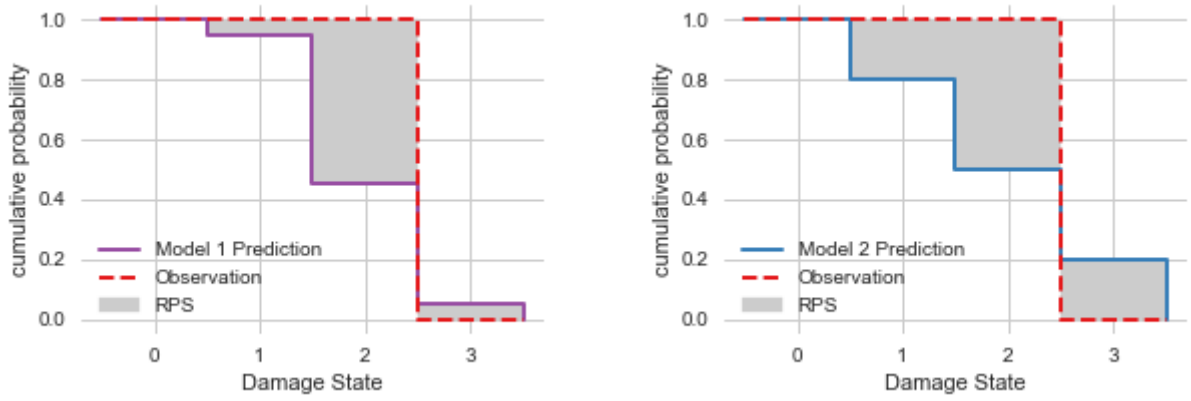
The rank probability score (RPS) is a metric used to evaluate the performance of a categorical probabilistic prediction. It is commonly used in weather forecasting to compare forecasts from different ensemble models [2]. The RPS computes the squared difference between the reverse CDF of the model prediction and the corresponding observation over a discrete set of prediction categories. The RPS is defined as

$$RPS = \sum_{k=1}^{K} (Y_k - O_k)^2 \tag{3}$$

where K is the number of damage states, Y_k are the kth components of the reverse CDF of the prediction (as defined in Eq. (2)), and O_k are the kth components of the reverse CDF of the observation. The RPS defined here uses the reverse CDF as opposed to the standard definition of RPS which uses the CDF. This is because the reverse CDF is able to capture the difference in probabilities of the highest damage state but not the lowest state (as evidenced in the pictorial representation in Fig. 1), while a CDF is the opposite and is able to capture the difference in the lowest state.

For the example in Table 1, Y is defined for models 1 and 2 in the previous section, and the observed damage state is 2, so $O = [1,1,1,0]$. Therefore, the RPS for model 1 is 0.31 and for model 2 is 0.35. RPS is a measure of error, so a smaller RPS is better, indicating that model 1 has better performance than model 2. This is a different conclusion than was reached using the expected and threshold damage accuracy metrics.

Fig. 1 shows a graphical representation of the RPS calculation for models 1 and 2, demonstrating that RPS can be interpreted as the areal difference between the observed and predicted reverse CDF. The RPS is a more general version of the Brier Score, and it extends the Brier Score to the multiclass and ordinal case [2].



(a) (b)

Fig. 1 – Example calculation of the RPS for (a) model 1 and (b) model 2 in Table 1.

The RPS directly compares the predicted probability distribution with the observed damage. There is no need to take the expectation or select a threshold. The RPS also takes into account the shape of the predicted distribution and accounts for the ordinality of the predicted categories. Damage states are ordinal in the sense that damage 0 is less than damage state 1, which is less than damage state 2, which is less than damage state 3. The RPS takes this ordinality into account because it uses the CDF for calculation of error. If one model predicts more probability in a damage state closer to the truth than another, that model will have a lower RPS. In this way, the RPS also accounts for the shape of the predicted distribution. A distribution with a more desirable unimodal shape will have a better RPS than a distribution with a less desirable bimodal shape.

One potential shortcoming of the RPS is that it does not consider whether the model over- or under-predicts the truth. Whether the probability outside the true category is assigned to a higher or lower damage state will not change the RPS. This might not be desirable behavior in certain cases, such as emergency management when it would be preferable to over- rather than under-predict damage.

Another disadvantage of RPS is that small perturbations in predictions of extreme events might not be captured well. One reason is that evaluating the model performance on rare occurrences by computing RPS only on observations of damage state 3 leads to an improper scoring rule and could lead to incorrect conclusions regarding relative model performance [7], [8].

2.4 Threshold-weighted Rank Probability Score (t-RPS)

The t-RPS is a modification on the RPS that puts weight on regions of the probability distribution of interest. For example, this can be used to put emphasis on matching the tails of the distribution for the prediction of rare events, such as severe damage from earthquakes.

The t-RPS for a discrete distribution is defined as

$$t - RPS = \sum_{k=1}^{K} w_k (Y_k - O_k)^2 \tag{4}$$

where K, Y_k, and O_k are the same as defined for RPS, and w_k is a weight specific to each damage state. The t-RPS has been described for a continuous distribution [8], [9], but this paper extends it to the discrete case.

For the example in Table 1, Y and O are defined in the previous section. If we assign weights, $w = [1, 10, 100, 1000]$, then t-RPS for model 1 is 32.8 and for model 2 is 65.6. Like RPS, t-RPS is a measure of error so a smaller score is better, indicating that model 1 has better performance than model 2. This is the same conclusion as reached by RPS, however the magnitude of the difference in scores is very different. The t-RPS for model 2 is twice the t-RPS for model 1, while the RPS for model 2 is only 13% larger than model 1, indicating that model 1 has better performance than model 2 for the higher damage states.

The t-RPS has the same advantages as the RPS: it allows for the direct comparison of a predicted probability distribution with observed damage, it considers the ordinality of the predicted categories, and it is a proper scoring rule [8], [9]. The t-RPS has the additional advantage that for two models with similar differences in the predicted probability but at different damage states, the model with a smaller difference at the damage state with higher weight will have a lower t-RPS. Therefore, in this example, the model with the better fit at higher damage states will be preferred over the other. It can also be used to emphasize small perturbations in the tails of the distribution, which is important for the prediction of rare events, like damage from earthquakes.

3. Simulation Study

Here we present the results of a simulation study in order to better understand the behavior of each metric described in the previous section. We define a simulation case study in which the true distribution of each observation is known. We obtain the observations (i.e., the ground truth labels), by taking random samples from the true probability distribution. We then understand the behavior of each metric by randomly perturbing the true distribution. In each experiment, 1 million samples were drawn.

For the first three experiments, we perturb the probability of the highest damage state by a varying amount for a binary classification problem and a balanced and imbalanced multiclass classification problem. We also include a fourth experiment in which we perturb the probability of each damage state by the same amount for a balanced and imbalanced multiclass classification problem. The imbalanced multiclass problem is the most similar to earthquake damage prediction because there are multiple damage states and the higher damage states tend to occur much less frequently than the lower damage states. For this reason, we use the class balanced [5] damage accuracy for all experiments.

3.1 Perturbation on highest damage state for balanced binary classification

In this experiment, we create a dataset that has two damage states, 0 and 1, and the probability of each damage state is equal. We generate model predictions by perturbing the true probability of damage state 1. This perturbation is applied by adding a random number to the probability of damage state 1, then subtracting that number from the probability of damage state 0 so that the total probability sums to 1. The maximum value of the perturbation is varied from -0.05 to +0.05. The perturbed probability distribution is then used to calculate each metric described in the previous section. For threshold balanced accuracy, the threshold is 0.5, and for t-RPS, the weights are [1, 10]. Results for each metric are shown in Fig. 2.

Because this is a binary classification example and the threshold was set to 0.5, the expected and threshold balanced accuracy are equal. The t-RPS is exactly 10 times the RPS because the weight on damage

state 1 was set to 10. Smaller values of RPS and t-RPS indicate better model performance, while larger values of expected and threshold accuracy indicate better performance. Since the minimum RPS and t-RPS occur at a perturbation of 0, these metrics correctly indicate that the best model prediction is the true distribution. The maximum expected and threshold accuracy occur at a perturbation of 0.01, however this can be attributed to slight variations in random sample generation.

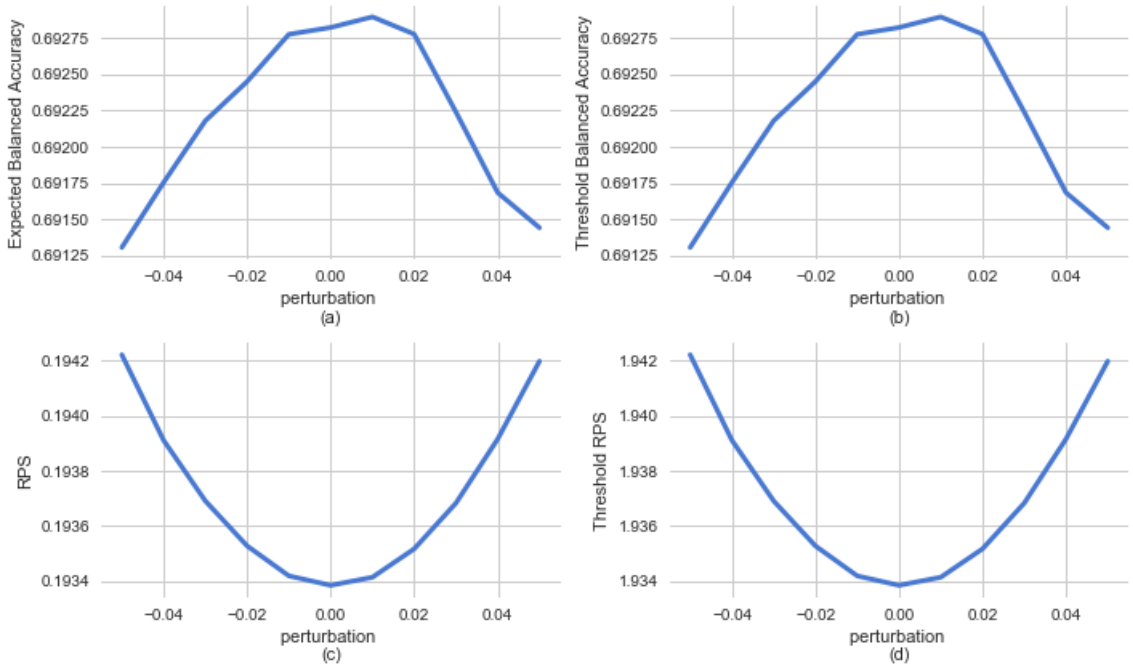


Fig. 2 – Value of the (a) expected balanced accuracy, (b) threshold balanced accuracy, (c) RPS, and (d) t-RPS for a balanced binary classification problem when the true probability of damage state 1 is perturbed by varying amounts.

3.2 Perturbation on highest damage state for balanced multiclass classification

In this experiment, we create a dataset that has four damage states, 0, 1, 2, and 3, and the probability of observing each damage state is equal. We generate model predictions by perturbing the true probability of damage state 3. This perturbation is applied by adding a random number to the probability of damage state 3, then subtracting that number from the probability of damage state 0 so that the total probability sums to 1. The maximum value of the perturbation is varied from -0.05 to +0.05. The perturbed probability distribution is then used to calculate each metric described in the previous section. For threshold balanced accuracy, the threshold is 0.5, and for t-RPS, the weights are [1, 10, 100, 1000]. Results for each metric are shown in Fig. 3.

The RPS and t-RPS have similar behavior for this multiclass example as the for the binary example shown in Fig. 2. Both metrics correctly indicate that the best model is the one without any perturbation on the true probability distribution, and the metrics indicate that performance decreases as the absolute value of the perturbation increases.

However, the maximum expected and threshold accuracy occur at perturbations other than 0, and their behavior does not follow any consistent trend that might be attributed to the natural variation in random sample generation. Hence, these metrics are not able to identify the true probability distribution as the best model.

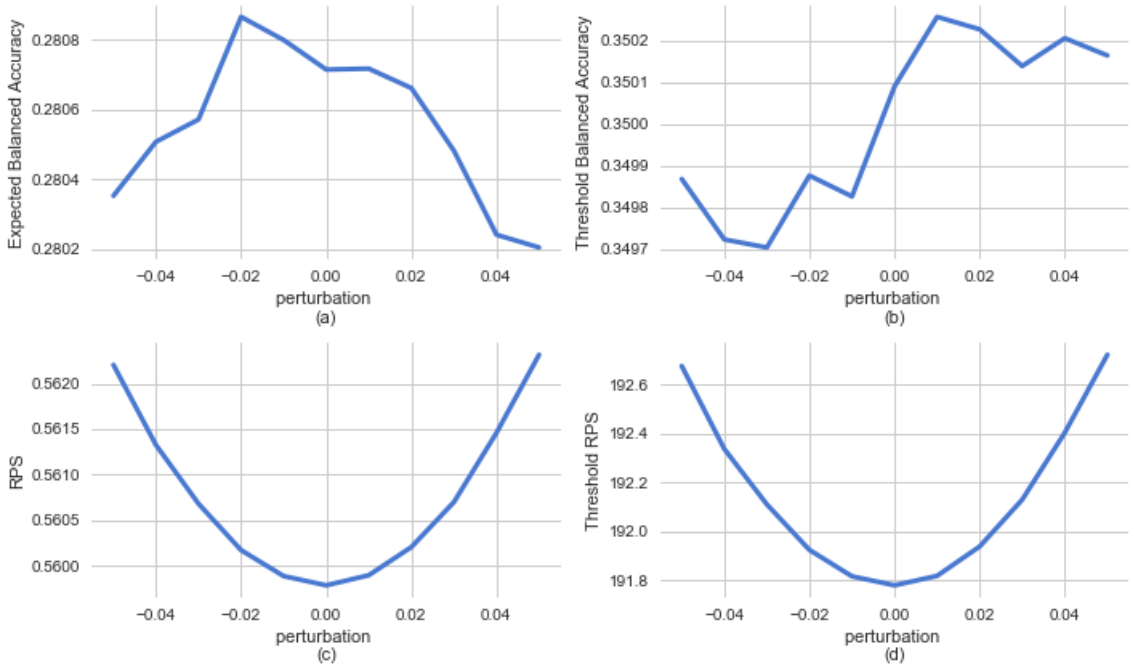


Fig. 3 – Value of the (a) expected balanced accuracy, (b) threshold balanced accuracy, (c) RPS, and (d) t-RPS for a balanced multiclass classification problem when the true probability of the highest damage state is perturbed by varying amounts.

3.3 Perturbation on highest damage state for imbalanced multiclass classification

In this experiment, we create a dataset that has four damage states, 0, 1, 2, and 3, and the probability of observing each damage state is imbalanced: 55% of the labels are damage state 0, 27% are damage state 1, 15% are damage state 2, and 3% are damage state 3. We generate the model predictions and calculate the metrics using the same method as described in the previous experiment. Results for each metric are shown in Fig. 4.

For this imbalanced case, the expected and threshold balanced accuracy again show undesirable behavior: the accuracy increases as the perturbation increases and the accuracy remains nearly constant as the perturbation decreases. The expected and threshold accuracy tend to predict damage states that are near the center of the distribution, which is damage states 0 and 1 for this imbalanced case, causing these metrics to miss the high damage states of 2 and 3. When the class balanced accuracy is taken, the missed predictions of damage states 2 and 3 have a large impact on the final accuracy. But when the predicted probability of damage state 3 is perturbed above the truth, the expected and threshold labeling approaches are more likely to predict damage states 2 and 3, bringing up the overall balanced accuracy.

The RPS and t-RPS are minimized by a perturbation of 0, again correctly indicating that the best model is the true probability of damage. For both the RPS and t-RPS, the score indicates relatively worse performance when the probability of damage state 3 is increased rather than decreased by the same amount. This is because in the imbalanced case, the true probability of damage state 3 is relatively small and can only be perturbed in the negative direction a small amount before reaching 0. The positive perturbation can be much larger before the probability of damage state 3 reaches 1. Therefore, the negative perturbation tends to be smaller than the positive perturbation for the imbalanced case, even when the maximum perturbation is the same.

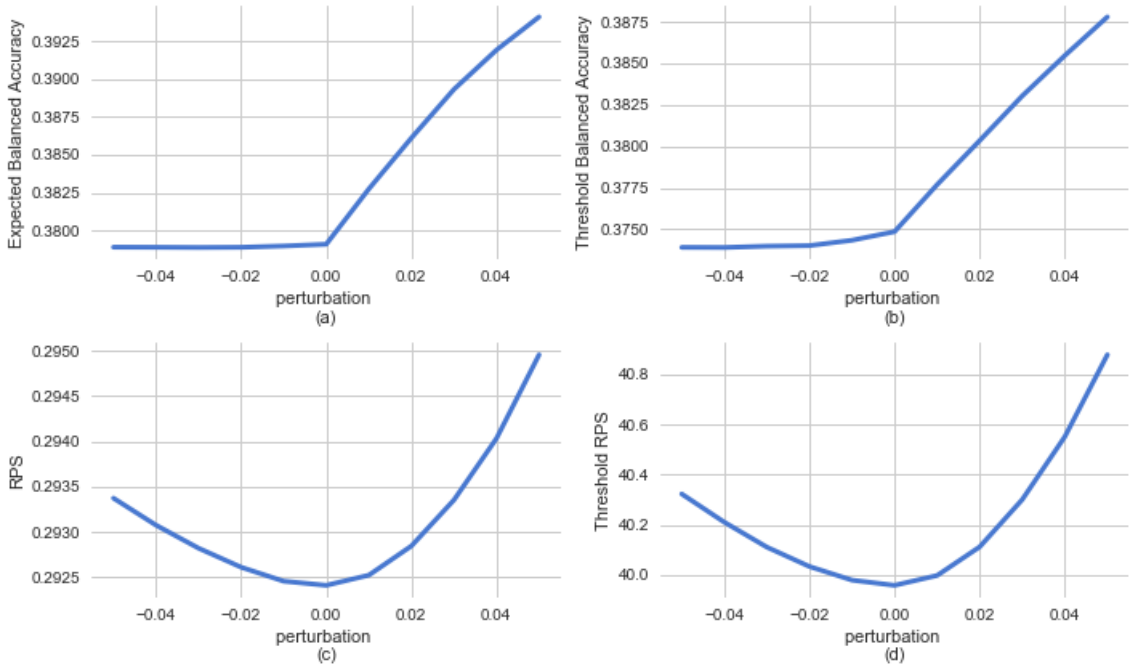


Fig. 4 – Value of the (a) expected balanced accuracy, (b) threshold balanced accuracy, (c) RPS, and (d) t-RPS for an unbalanced multiclass classification problem when the true probability of the highest damage state is perturbed by varying amounts.

3.4 Perturbation on each damage state for multiclass classification

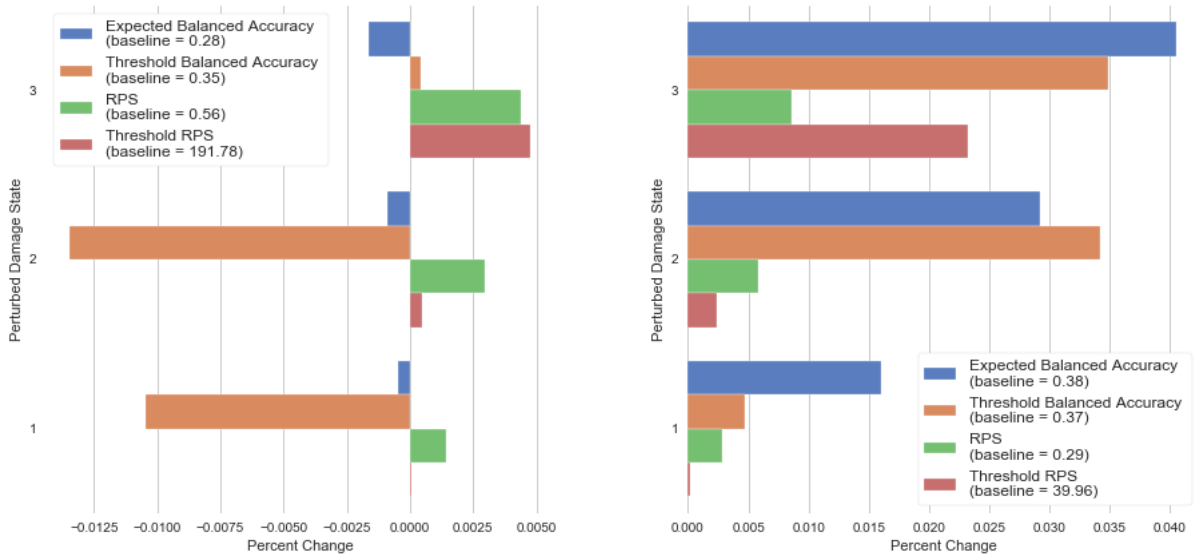
In this experiment, we create a dataset that has four damage states, 0, 1, 2, and 3, and the probability of observing each damage state is equal. We create the true probabilities as described in previous experiments, and we generate the model predictions by perturbing the predicted probability of each damage state by a random number with a maximum of +0.05, and always balancing the perturbation by changing the probability of damage state 0 to keep the sum to 1. For each metric and each perturbed damage state, we compute the percent difference between the metric computed on the true distribution and the metric computed on the perturbed distribution. Results are shown in Fig. 5a.

We also perform the same experiment for a dataset where the probability of observing each damage state is imbalanced, similar to Section 3.3. Results for this case are shown in Fig. 5b. In both cases, the class balanced expected and threshold damage accuracy is computed. A threshold of 0.5 is used for the threshold damage accuracy and weights of [1, 10, 100, 1000] are used for t-RPS.

For the balanced case, the expected accuracy decreases for a perturbation applied at each damage state, which is the expected behavior. However, the percent of the decrease is relatively small compared to other metrics, especially for damage state 3. The threshold accuracy decreases significantly for perturbations applied at damage states 1 and 2 but increases for a perturbation applied at damage state 3, incorrectly indicating that a model with perturbations on damage state 3 is a better model than the true distribution. For the imbalanced case, both the expected and threshold accuracy increase as perturbations are applied at any damage state, giving further evidence that these metrics are not suitable for the evaluation of a probabilistic model because they are not maximized by the model with the true distribution.

For both the balanced and imbalanced cases and across perturbations on all damage states, the RPS and t-RPS increase in value, indicating reduced model performance. A perturbation applied at damage state 3 causes a larger percent increase in RPS and t-RPS than the same perturbation applied at damage states 2 or 1. This is because both metrics evaluate the error of the reverse CDF (defined in Eq. (2)), so perturbations in damage state 3 affect damage states 1 and 2, while perturbations in damage state 1 do not affect damage states

2 or 3. A perturbation at damage state 3 causes roughly a 3 times increase in the RPS for the same perturbation at damage state 1. This effect is magnified by the t-RPS because of the additional weight applied on the error in the reverse CDF at each damage state.



(a) (b)

Fig. 5 – Percent change in each metric when damage states 1, 2, and 3 are perturbed by a maximum value of +0.05 for (a) balanced and (b) imbalanced multiclass classification problems.

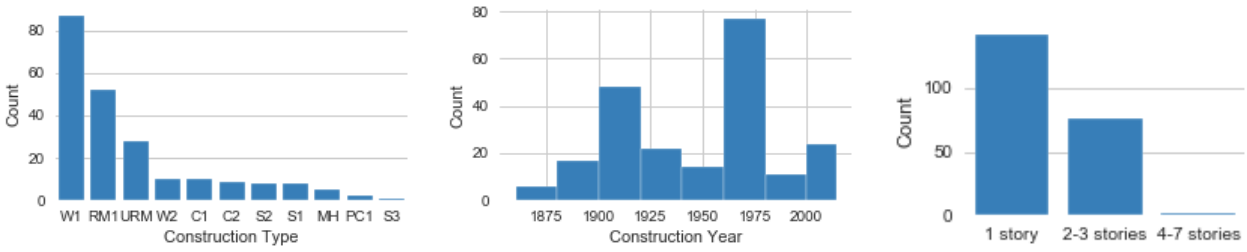
4. Case Study

Here we present a case study using actual building damage data from the 2014 earthquake in Napa, California, USA. Damage data from the Napa earthquake is publicly available from the EERI Earthquake Clearinghouse for about 92,000 buildings [10]. The data is labeled in the form of ATC-20 tags [11], i.e., green, yellow, and red, which we map to damage states 1, 2, and 3, respectively. However, only 220 of those buildings include the construction type, construction year, and number of stories, which is enough detail to assign a Hazus fragility function. The Hazus software was developed by the United States (US) Federal Emergency Management Agency (FEMA) and is used extensively in the US for earthquake damage estimation [12]. 220 buildings out of 92,000 is a small and likely non-representative sample, which will affect the performance of the metrics. It is important to also note that even the 92,000 buildings are likely not a representative sample of damage from the Napa earthquake because field reconnaissance data tends to be biased towards damaged buildings. Data about undamaged buildings is rarely collected and therefore undamaged buildings are often underrepresented with respect to their true distribution in these types of datasets.

We assign peak ground acceleration (PGA) to each building using cubic interpolation from the USGS ShakeMap for the 2014 Napa earthquake [13]. USGS ShakeMaps provide a map of estimated shaking intensity measures, including PGA, shortly after an earthquake [14]. Using PGA from the USGS ShakeMap and fragility functions from Hazus, we estimate the probability of each damage state. Hazus defines damage states as none, slight, moderate, extensive, and complete. We map the none damage state to damage 0, slight to damage state 1, moderate to damage state 2, and extensive and complete to damage state 3.

The performance of the Hazus probabilistic predictions are compared to a simple baseline model. The baseline model uses the potential damage scale defined by the USGS ShakeMap to assign a damage state [13]. The baseline model assumes that PGA values between 0 and 6.2%g are damage state 0, 6.2 to 22%g are damage state 1, 22 to 40%g are damage state 2, and >40%g is damage state 3. The probabilistic prediction from this model assumes a probability of 1 for the assigned damage state and 0 for all others.

See Fig. 6 for a summary of the building characteristics in the damage dataset. Most buildings are 1-story wood or reinforced masonry, constructed around 1970.



(a) (b) (c)

Fig. 6 – Distribution of (a) construction types, (b) construction year, and (c) number of stories of damaged buildings in the 2014 Napa earthquake.

Table 2 shows the number of buildings in each damage state for the ground truth, and the predictions from the baseline and Hazus models. The number of buildings in each damage state for the Hazus prediction is computed by summing the probability of each damage state across the entire dataset. The baseline model predicts that almost all buildings are in damage state 2, while the ground truth and Hazus are more evenly distributed across damage states.

Table 2 – Number of buildings in each damage state in the 2014 Napa damage dataset, observed and predicted by the baseline and Hazus models.

Damage State	Truth	Baseline	Hazus
0	97	0	53.4
1	52	2	49.4
2	36	217	53.9
3	35	1	63.3

Table 3 shows each of the metrics computed for the baseline and Hazus models. The expected and threshold damage accuracy are class balanced. A threshold of 0.5 was used for threshold accuracy and weights of [1, 10, 100, 1000] were used for t-RPS. All metrics indicate that the Hazus model has better performance than the baseline. However, the magnitude of the difference in performance is relatively small for the expected and threshold accuracy, while it is much larger for RPS and t-RPS.

Table 3 – Metrics computed for the predicted damage in the 2014 Napa earthquake using the Hazus and baseline models.

Model	Expected Accuracy	Threshold Accuracy	RPS	t-RPS
Hazus	0.295	0.305	0.713	222.8
Baseline	0.248	0.248	1.282	235.8

In order to understand the variation of each metric, bootstrapping was used on the damage dataset. We randomly sampled 220 buildings with replacement from the original dataset of 220 buildings. This was repeated 10,000 times and the metrics were computed for each bootstrapped dataset. The distribution of each metric for the baseline and Hazus models is shown in Fig. 7. The RPS is the only metric for which the Hazus model is always considered to be the best model, regardless of the distribution of the underlying data. The t-RPS is very sensitive to the distribution of the underlying data because damage state 3 is rare and the weight on damage state 3 is large. This indicates that t-RPS should be used carefully and only when the distribution of the damage data is known to be a representation sample of the overall damage distribution.

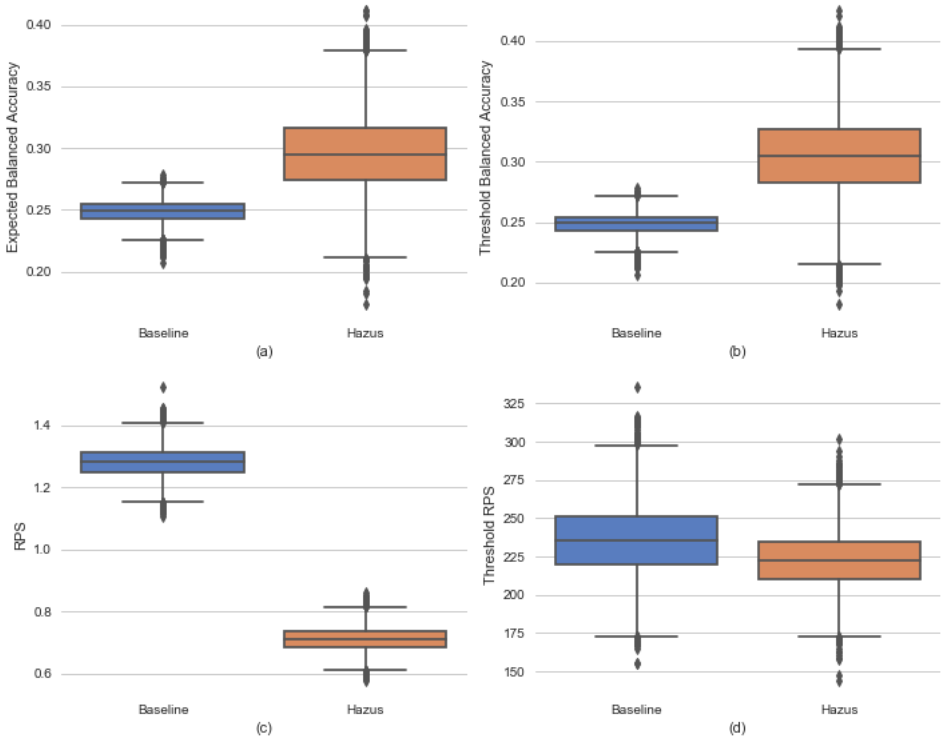


Fig. 7 – Distribution of (a) expected balanced accuracy, (b) threshold balanced accuracy, (c) RPS, and (d) t-RPS after performing bootstrapping on the 2014 Napa damage dataset.

5. Conclusions

We investigated four metrics for the evaluation of probabilistic models through a simulation study and a case study of actual damage from the 2014 Napa earthquake. The four metrics are: (1) expected damage accuracy, (2) threshold damage accuracy, (3) rank probability score (RPS), and (4) threshold-weighted rank probability score (t-RPS).

Expected and threshold damage accuracy are simple to compute and communicate, but we demonstrated that they are not suitable for the evaluation of probabilistic models because they are not maximized by the true distribution. In particular, for an imbalanced multiclass problem like earthquake damage prediction, the expected and threshold class balanced accuracy indicate that a model with a large positive perturbation on damage state 3 has better performance than the true distribution. This could lead to a model that significantly overpredicts damage, potentially making it difficult for emergency personnel to correctly prioritize areas that require immediate assistance.

RPS and t-RPS are shown to be suitable for probabilistic model evaluation because they are minimized by the true distribution. This is true for binary, multiclass, balanced, and imbalanced classification problems. These metrics also account for ordinality in the prediction categories. Because t-RPS with the weights proposed here puts higher importance on the tails of the probability distribution, it is more sensitive to the observed damage distribution in the underlying data. This is desirable when the distribution of the observed damage is known to be representative of the overall damage data. In this case, t-RPS encourages the correct prediction of the higher damage states on which a larger weight is placed.

In general, these metrics do not solve the problem that earthquake damage data tends to be biased toward damage. This is because data about undamaged buildings is rarely collected, so damaged buildings tend to be overrepresented in building damage datasets. It may not be possible to capture the true probability distribution with a non-random sample [15]. Therefore, we suggest the use of these metrics only in cases where the true probability of damage can be reasonably estimated.

7. References

[1] R. Dinga, B. W. J. H. Penninx, D. J. Veltman, L. Schmaal, and A. F. Marquand, "Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines," Neuroscience, preprint, Aug. 2019.
[2] A. P. Weigel, M. A. Liniger, and C. Appenzeller, "The Discrete Brier and Ranked Probability Skill Scores," *Mon. Weather Rev.*, vol. 135, no. 1, pp. 118–124, Jan. 2007, doi: 10.1175/MWR3280.1.
[3] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951, doi: 10.1214/aoms/1177729694.
[4] C. Beckham and C. Pal, "Unimodal probability distributions for deep ordinal classification," *ArXiv170505278 Stat*, Jun. 2017.
[5] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124, doi: 10.1109/ICPR.2010.764.
[6] F. Atger, "Verification of intense precipitation forecasts from single models and ensemble prediction systems," *Nonlinear Process. Geophys.*, vol. 8, no. 6, pp. 401–417, 2001.
[7] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: 10.1198/016214506000001437.
[8] S. Lerch, "Verification of probabilistic forecasts for rare and extreme events," Heidelberg University, Heidelberg, Germany, 2012.
[9] T. Gneiting and R. Ranjan, "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules," *J. Bus. Econ. Stat.*, vol. 29, no. 3, pp. 411–422, Jul. 2011, doi: 10.1198/jbes.2010.08110.
[10] Earthquake Engineering Research Institute (EERI), "South Napa, USA Earthquake Clearinghouse," 2014. [Online]. Available: http://www.eqclearinghouse.org/2014-08-24-south-napa/.
[11] Applied Technology Council, "Procedures for Post-Earthquake Safety Evaluation of Buildings," Redwood City, CA, ATC-20, 1989.
[12] "Hazus MH 2.1 Earthquake Model Technical Manual," Department of Homeland Security, Federal Emergency Management Agency (FEMA), Mitigation Division, Washington, D.C., 2013.
[13] US Geological Survey (USGS), "M 6.0 - South Napa," *USGS Earthquake Hazards Program*, Aug-2014. [Online]. Available: https://earthquake.usgs.gov/earthquakes/eventpage/nc72282711/executive.
[14] D. J. Wald, B. C. Worden, V. Quitoriano, and K. L. Pankow, "ShakeMap Manual: Technical Manual, Users Guide, and Software Guide," US Geological Survey, Boulder, CO, 1.0, 2006.
[15] J. B. Copas and H. G. Li, "Inference for Non-random Samples," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 59, no. 1, pp. 55–95, Feb. 1997, doi: 10.1111/1467-9868.00055.

